学修番号 19860634

修士論文

文法誤り訂正における
訂正度を考慮した多様な訂正文の生成

甫立 健悟

2021 年 2 月 19 日

東京都立大学大学院
システムデザイン研究科 情報科学域

甫立 健悟

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

文法誤り訂正における
訂正度を考慮した多様な訂正文の生成*

甫立 健悟

修論要旨

文法誤り訂正は言語学習者の書いた文法的に誤りを含んだ文を文法的に正しい文へと訂正を行うタスクであり，第二言語学習者の作文支援システムとして有用である．文法的に誤りを含んだ文に対して訂正を行う際，訂正結果は複数存在することがある．例えば，Bryant and Ng は，文法的に誤りを含んだ文に対して 10 人のアノテータがそれぞれ異なる有効な訂正手法を提案する場合があることを示した．この 10 人のアノテータに対して，明示的に異なる訂正文の作成を行わせていないが，実際に，訂正した文におけるアノテータ間の一致率は約 16% であった．この様に，様々な訂正結果が存在するため，文法誤り訂正モデルが複数の訂正結果を提示することで，言語学習者は訂正結果を反映するかどうかの判断や，複数の訂正結果の中から好みの表現を選択することが可能になる．

一般に，文法誤り訂正は文法的に誤りを含んだ文から文法的に正しい文への機械翻訳タスクとして捉えられ，近年，ニューラルネットワークを用いた機械翻訳モデルが文法誤り訂正モデルとして用いられることが多い．実際に，機械翻訳モデルを文法誤り訂正モデルに適用することにより，文法誤り訂正タスクにおいても有効な結果を示している．しかしながら，既存の文法誤り訂正モデルは 1 つの入力文に対して 1 つの有効な出力文の生成を目指しており，複数の訂正結果の生成を考慮していない．そこで本研究では，文法誤り訂正において多様な出力を生成するという新たなタスクに取り組み，訂正度を制御可能な文法誤り訂正モデルを用いた手法を提案する．

訂正を行う際，必要最低限の書き換えのみ行うか，もしくは，より多くの書き換えを行うかという違いにより，訂正文に多様さが生じる．Sakaguchi らは，専門性

*東京都立大学大学院 システムデザイン研究科 情報科学域 修士論文, 学修番号 19860634,
2021 年 2 月 19 日.

の異なるアノテータにおいて，必要最小限の書き換えと流暢性を求めた書き換えの 2 種類の訂正をそれぞれ行うことで，多様な訂正文を作成し，それらの一致率は約 15% であったと報告している．つまり，1 つの誤り文に対しても複数の訂正度を用いて訂正が行われるということである．しかし，既存の文法誤り訂正モデルは学習した単一の訂正度でのみ訂正を行っており，それらの異なる訂正度で訂正を行う手法の研究は行われていない．そこで本研究では，訂正度を制御可能な文法誤り訂正モデルを提案し，1 つの入力文に対して複数の訂正度での訂正文を生成することで，多様な訂正文の生成を行う．手法としては，まず，文法誤りが訂正されているデータ内において，1 文ごとの訂正度の情報を特殊トークンとして文に付与し，新たな訓練データを作成する．そして，新たに作成した訓練データを用いて文法誤り訂正モデルの学習を行い，推論時には，入力文に任意の訂正度の特殊トークンを付与することで，モデルの訂正度を制御する．結果として，1 つの入力文に対して 1 つの訂正度ではなく，複数の訂正度を用いて訂正を行うことが可能となり，多様な訂正文を得ることが可能となる．

本研究では，さらに出力を多様化する手法として，誤り箇所を考慮したビームサーチ手法を提案する．文法誤り訂正において，複数の訂正結果を生成する手法としてはビームサーチを用いて n-best を生成する方法が存在する．しかし，これらの研究では，1 つの適切な訂正文の出力の探索のためにビームサーチを利用しており，多様な訂正結果の出力を目的としていない．さらに，通常のビームサーチを用いた n-best 出力は多様性に欠けることが示されている．そのため，機械翻訳の分野などにおいて，多様な候補を生成するために，出力を多様にする制約を加えたいくつかの多様なビームサーチ手法が提案されている．これらの多様なビームサーチ手法は，文中の全てのトークンに対して様々な書き換えを行うことで，複数の出力文に多様性をもたらしている．一方で，文法誤り訂正の様な入力文と出力文の大部分が共通しているようなタスクにおいては，これらの手法は適していないと考えられる．そこで本研究では，誤り箇所を考慮したビームサーチ手法を提案する．誤り箇所を考慮したビームサーチでは，訂正が必要な単語に対してのみ様々なパスの探索を行う．それゆえ，提案手法では，訂正が必要な箇所でのみ，通常のビームサーチよりも多様な単語を組み合わせた文の生成を行うことが可能となる．

実験の結果，訂正度を制御可能な文法誤り訂正モデルを用いることで既存手法よ

りも多様な訂正結果を生成することが可能となり，誤り箇所を考慮したビームサーチと組み合わせることで，更に多様化可能であることを示した．

本研究の主な貢献は以下の 4 つである．

- 単語編集率により文法誤り訂正モデルの訂正度が制御可能なことを示した．
- 既存の多様な文を生成する手法が文法誤り訂正においては適切な多様性をもたらさないことを示した．
- 訂正度を制御した文法誤り訂正モデルの出力を用いることで多様な出力が得られることを示した．
- 誤り箇所を考慮したビームサーチを提案し，訂正度を制御した文法誤り訂正モデルと組み合わせることで既存手法よりも適切に出力文に多様性をもたらすことを示した．

本稿の構成は以下の通りである．第 1 章では，本研究の提案，貢献，概要について述べる．第 2 章では，既存の出力文の制御や多様化の先行研究について紹介する．第 3 章では，訂正度を制御した文法誤り訂正モデルを提案する．第 4 章では，誤り箇所を考慮したビームサーチを提案する．第 5 章では，BERT を用いた文分類を提案する．第 6 章では，複数の人手による訂正文が付与されている評価データを用いて提案手法を評価する．第 7 章では，提案モデルについて分析する．最後に第 8 章で，本研究のまとめを述べる．

Generation of Diverse Corrected Sentences Considering the Degree of Correction*

Kengo Hotate

Abstract

Grammatic error correction (GEC) is a task to correct a sentence containing grammatical errors written by language learners into a grammatically correct sentence, and is useful as a writing support system for second language learners. Depending on the input, there are multiple ways to correct such text. For example, Bryant and Ng showed that 10 annotators can produce 10 different valid correction results for the same grammatically incorrect text. If a GEC model presents multiple candidates for correction, it helps the user decide whether to utilize the correction results such that the user can select a favorite correct expression from among the candidates.

In general, GEC is regarded as a machine translation task from grammatically incorrect sentences to grammatically correct sentences. Therefore, in recent years, machine translation models based on neural networks are often used as GEC models. In fact, the adaptation of machine translation models to GEC models has shown effective results in the GEC task. However, existing GEC models aim to produce a single valid output sentence for a single input sentence, and do not consider the generation of multiple correction results. Therefore, we address a new task of generating diverse outputs in GEC, and propose a method using a GEC model that can control the degree of correction.

When making corrections, there are diverse corrected sentences depending on whether making only the minimum edits or making more edits. Sakaguchi et

*Master's Thesis, Department of Computer Systems, Graduate School of System Design, Tokyo Metropolitan University, Student ID 19860634, February 19, 2021.

al. reported that a diverse of corrective sentences were produced for annotators with different specialties by making two types of corrections, one for minimal rewriting and the other for fluency, and that the agreement rate between them was about 15%. In other words, even one error sentence is corrected using multiple correction degrees. However, existing GEC models correct only with a single learned degree of correction, and there is no research on methods that correct with these different degrees of correction. In this thesis, we propose a GEC model that can control the degree of correction, and generate multiple degrees of correction for a single input sentence to generate diverse corrected sentences. First, in the data where grammatical errors are corrected, we create new training data by assigning information on the degree of correction for each sentence as a special token to the sentence. Second, we train the GEC model using the newly created training data, and during inference, we assign special tokens of arbitrary degree of correction to the input sentences to control the degree of correction of the model. As a result, it is possible to make corrections using multiple degrees of correction instead of just one for a single input sentence, thus obtaining diverse corrected sentences.

In this thesis, we propose a beam search method that considers the location of errors to diversify the output further. There are several methods for generating multiple correction results in GEC, such as generating n-best using beam search. However, in these studies, beam search is used to find the output of a single appropriate correction sentence, and not for the purpose of outputting diverse correction results. Furthermore, it has been shown that the n-best output using an plain beam search lacks diversity. Therefore, in machine translation and other fields, several diverse beam search methods have been proposed to generate diverse candidates, with the addition of constraints to make the output diverse. These diverse beam search methods provide diversity in multiple output sentences by performing diverse rewritings on all the tokens in the sentence. On the other hand, these methods are not suitable for tasks where most of the input and output sentences are common. In this thesis, we propose a beam search

method that considers the error points. Our proposed method, the diverse local beam search, searches for various paths only for words that need to be corrected. Therefore, our proposed method can generate sentences that combine more diverse words than plain beam search only for the words that need to be corrected.

As a result of experiments, we found that a GEC model that can control the degree of correction can produce more diverse correction results than existing methods. We also showed that it is possible to further diversify by combining it with the diverse local beam search.

The main contributions of this work are summarized as follows:

- We showed that the degree of correction in a GEC model can be controlled by the WER.
- We showed that existing methods for generating diverse sentences do not provide adequate diversity in GEC.
- We showed that using a GEC model with a controlled degree of correction can produce diverse outputs.
- We proposed a beam search that considers error points and showed that combining it with a GEC model controls the degree of correction provides more diversity to the output sentences than the existing method.

This thesis comprises as follows. In Chapter 1, we introduce an overview and contributions of our work. In Chapter 2, we introduce the existing previous work on GEC, output sentence control, and diversification. In Chapter 3, we propose a GEC model with a controlled degree of correction. In Chapter 4, we propose a beam search method that considers the error points. In Chapter 5, we propose a sentence classification method using BERT. In Chapter 6, we evaluate the proposed method using evaluation data with multiple manually corrected sentences. In Chapter 7, we analyze the proposed model. Finally, in Chapter 8, we give a summary of this thesis.

目次

図目次

原文	**To put it in the nutshell**, I believe that people should **have the obligation** to tell their relatives about **the** genetic **testing result** for the good of their health.
	To put it in **a** nutshell, I believe that people should **be obliged** to tell their relatives about **their** genetic **test results** for the good of their health.
	In a nutshell, I believe that people should have **an** obligation to tell their relatives about the genetic testing result for the good of their health.
参照文	**In summary**, I believe that people should have the obligation to tell their relatives about the genetic testing result for the good of their health.
	In a nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing result for the good of their health.
	To put it in **a** nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing **results** for the good of their health.

表 1.1: Bryant and Ng (2015) により作成された 1 文に対する複数の訂正文の例

第 1 章　はじめに

文法誤り訂正は言語学習者の書いた文法的に誤りを含んだ文を文法的に正しい文へと訂正を行うタスクであり，第二言語学習者の作文支援システムとして有用である．文法的に誤りを含んだ文に対して訂正を行う際，訂正結果は複数存在することがある．例えば，Bryant and Ng [1] は，文法的に誤りを含んだ文に対して 10 人のアノテータがそれぞれ異なる有効な訂正手法を提案する場合があることを示した．この 10 人のアノテータに対して，明示的に異なる訂正文の作成を行わせていないが，実際に，訂正した文におけるアノテータ間の一致率は約 16% であった．表 1.1 に Bryant and Ng [1] により作成された複数訂正文の例の一部を示す．**太字**は原文から訂正が行われた箇所を示している．1 文から複数の訂正文が提案されているが，

いずれも原文において誤りのある箇所のみにおいて多様な訂正が行われている．この様に，様々な訂正結果が存在するため，文法誤り訂正モデルが複数の訂正結果を提示することで，言語学習者は訂正結果を反映するかどうかの判断や，複数の訂正結果の中から好みの表現を選択することが可能になる．

一般に，文法誤り訂正は文法的に誤りを含んだ文から文法的に正しい文への機械翻訳タスクとして捉えられ，近年，ニューラルネットワークを用いた機械翻訳モデルが文法誤り訂正モデルとして用いられることが多い．実際に，機械翻訳モデルを文法誤り訂正モデルに適用することにより，文法誤り訂正タスクにおいても有効な結果を示している [2, 3, 4, 5, 6]．しかしながら，既存の文法誤り訂正モデルは 1 つの入力文に対して 1 つの有効な出力文の生成を目指しており，複数の訂正結果の生成を考慮していない．そこで本研究では，文法誤り訂正において多様な出力を生成するという新たなタスクに取り組み，訂正度を制御可能な文法誤り訂正モデルを用いた手法を提案する．

訂正を行う際，必要最低限の書き換えのみ行うか，もしくは，より多くの書き換えを行うかという違いにより，訂正文に多様さが生じる．文法誤り訂正タスクにおいて訓練データとして一般的に使用される Lang-8 [7]，評価データとして使用される CoNLL-2014 [8] や JFLEG [9] では，1 文中における訂正の量という意味での訂正度が異なることが知られている．また，Sakaguchi ら [10] は，専門性の異なるアノテータにおいて，必要最小限の書き換えと流暢性を求めた書き換えの 2 種類の訂正をそれぞれ行うことで，多様な訂正文を作成し，それらの一致率は流暢な訂正において約 15% 以下，最小限な訂正においても約 38% 以下であったと報告している．つまり，1 つの誤り文に対しても複数の訂正度を用いて訂正が行われるということである．つまり，1 つの誤り文に対しても複数の訂正度を用いて訂正が行われるということである．しかし，既存の文法誤り訂正モデルは学習した単一の訂正度でのみ訂正を行っており，それらの異なる訂正度で訂正を行う手法の研究は行われていない．そこで本研究では，訂正度を制御可能な文法誤り訂正モデルを提案し，1 つの入力文に対して複数の訂正度での訂正文を生成することで，多様な訂正文の生成を行う．手法としては，まず，文法誤りが訂正されているデータ内において，1 文ごとの訂正度の情報を特殊トークンとして文に付与し，新たな訓練データを作成する．ここで，訂正度を表す指標として単語編集率を用いる．単語編集率とは文中

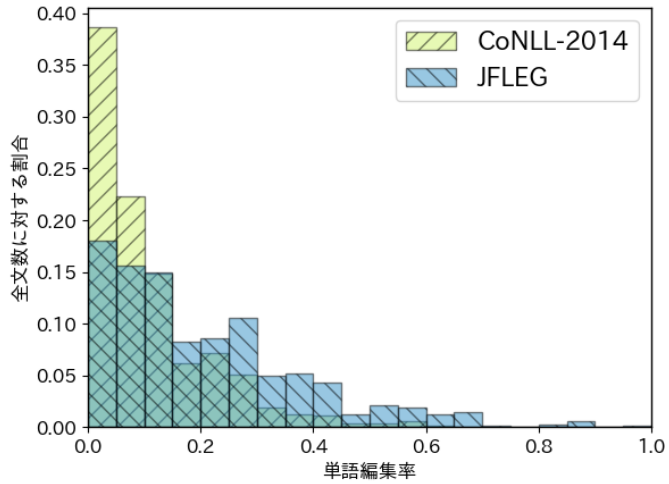

図 1.1: １文中の単語編集率のヒストグラム

の単語がどれだけ書き換えられたのかを表す指標であるため，文法的誤りを含んだ文と，その誤りを訂正した文の単語編集率は，文の訂正度を表していると言える．CoNLL-2014 と JFLEG では， JFLEG の方が訂正度が大きいことが知られており，実際に，図 1.1 に示すグラフからも，CoNLL-2014 よりも JFLEG の方が単語編集率が大きいため，単語編集率が訂正度を示していることが確認できる．そして，新たに作成した訓練データを用いて文法誤り訂正モデルの学習を行い，推論時には，入力文の文に任意の訂正度の特殊トークンを付与することで，付与した単語編集率に基づきモデルの訂正度を制御する．結果として，1 つの入力文に対して 1 つの訂正度ではなく，複数の訂正度を用いて訂正を行うことが可能となり，多様な訂正文を得ることができる．

本研究では，さらに出力を多様化する手法として，誤り箇所を考慮したビームサーチ手法を提案する．文法誤り訂正において，複数の訂正結果を生成する手法としてはビームサーチを用いて n-best を生成する方法が存在する [11, 6]．しかし，これらの研究では，1 つの適切な訂正文の出力の探索のためにビームサーチを利用しており，多様な訂正結果の出力を目的としていない．さらに，通常のビームサーチを用いた n-best 出力は多様性に欠けることが示されている [12]．そのため，機械翻訳の分野などにおいて，多様な候補を生成するために，出力を多様にする制約

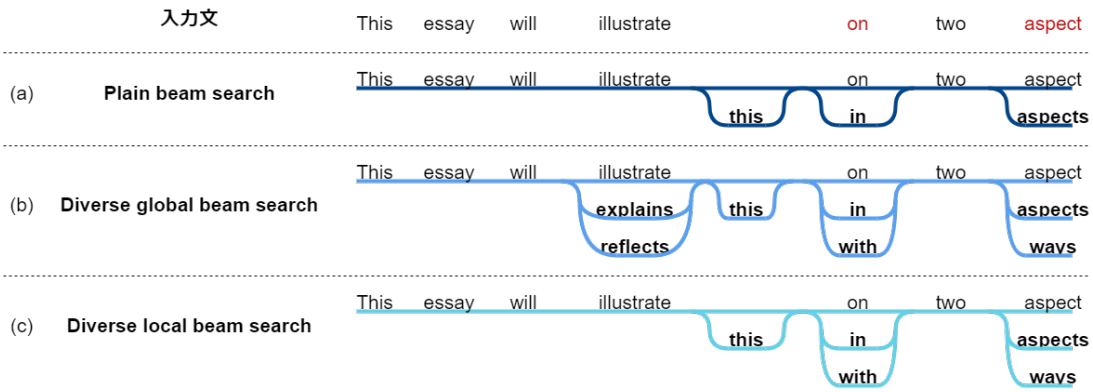

図 1.2: 既存のビームサーチ手法と提案手法を比較した概要図

を加えたいくつかの多様なビームサーチ手法が提案されている [12, 13]．これらの多様なビームサーチ手法は，文中の全てのトークンに対して様々な書き換えを行うことで，複数の出力文に多様性をもたらしている．一方で，入力文と出力文の大部分が共通しているようなタスクにおいては，これらの手法は適していないと考えられる．特に，文法誤り訂正においては，入力文の文法的に正しい部分に対しても書き換えが行われてしまうため，この様な入力文の全体に対して書き換えを行う手法は好ましくない．そのため，文法誤り訂正モデルは，入力文中の文法的に正しい部分は書き換えずに，誤りを含む部分に対してのみ様々な方法で訂正を行うことが望まれる．そこで本研究では，誤り箇所を考慮したビームサーチ手法を提案する．図 1.2 は既存手法と提案手法との比較を図示したものである．赤字の単語は誤りであることを，太字の単語は訂正が行われたことを示している．

(a) 通常のビームサーチ（Plain beam search）では，訂正が特定のパスに集中しているため多様性がなく，類似した単語の組み合わせで文を生成している．
(b) 既存手法の多様なビームサーチ（Diverse global beam search）は，様々なパスを探索している．したがって，通常のビームサーチとは異なり，この方法では様々な単語の組み合わせで文が生成されるため，多様な出力を得ることができる．ただし，訂正する必要のない単語においても書き換えを行った出力も生成されてしまう．
(c) 提案手法である誤り箇所を考慮したビームサーチ（Diverse local beam search）

では，訂正が必要な単語に対してのみ様々なパスの探索を行う．それゆえ，提案手法では，訂正が必要な箇所でのみ，通常のビームサーチよりも多様な単語を組み合わせた文の生成を行う

ここで，上記の手法は全て同じビーム幅であるが，パスが異なることに注意されたい．

実験の結果，訂正度を制御可能な文法誤り訂正モデルを用いることで既存手法よりも多様な訂正結果を生成することが可能となり，誤り箇所を考慮したビームサーチと組み合わせることで，更に多様化可能であることを示した．

本研究の主な貢献は以下の 4 つである．

- 単語編集率により文法誤り訂正モデルの訂正度が制御可能なことを示した．
- 既存の多様な文を生成する手法が文法誤り訂正においては適切な多様性をもたらさないことを示した．
- 訂正度を制御した文法誤り訂正モデルの出力を用いることで多様な出力が得られることを示した．
- 誤り箇所を考慮したビームサーチを提案し，訂正度を制御した文法誤り訂正モデルと組み合わせることで既存手法よりも適切に出力文に多様性をもたらすことを示した．

本稿の構成を示す．第 2 章では，既存の出力文の制御や多様化の先行研究について紹介する．第 3 章では，訂正度を制御した文法誤り訂正モデルを提案する．第 4 章では，誤り箇所を考慮したビームサーチを提案する．第 5 章では，BERT を用いた文分類を提案する．第 6 章では，複数の人手による訂正文が付与されている評価データを用いて提案手法を評価する．第 7 章では，提案モデルについて分析する．最後に第 8 章で，本研究のまとめを述べる．

第 2 章　関連研究

2.1　文法誤り訂正

文法誤り訂正は文法的に誤りを含んだ文から正しい文へと翻訳するタスクと捉えられることが多い．そのため，文法誤り訂正に関する研究は，主に機械翻訳において有効であると示された手法を用いることが多い．Chollampatt and Ng [2] は CNN を，Junczys-Dowmunt ら [3] は Transformer [14] を文法誤り訂正モデルに適応することで高い性能を発揮することを示した．本研究においても，文法誤り訂正モデルとして Transformer を用いた実験を行った．ここで，Transformer とは機械翻訳の分野において，自己注意機構を用いることで大幅な性能向上を達成したモデルである．さらに，CNN などと比較して計算量も小さくなっているため，大規模データを用いた学習にも適している．

Zhao ら [4] は Trasnformer にコピー機構を用いることで，精度向上を図った．さらに，Kiyono ら [5] は Trasnformer の学習に用いるデータを約 70M もの大規模な疑似データを用いることで，精度向上することを示した．本研究では，訂正性能の向上ではなく多様性の向上に焦点を当てているため，通常の Transformer を用い，疑似データは使用していない．

2.2　多様な出力

出力文に対して制御を行う研究は，文法誤り訂正の分野ではあまり行われておらず，主に機械翻訳の分野において行われている．Sennrich ら [15] は，機械翻訳において訓練データに文の丁寧さの情報を特殊トークンとして入力文に付与し，モデルの訓練を行うことで出力文の敬意表現の制御を行った．敬語の存在しない言語から敬語の存在する言語への翻訳時，訓練データの入力文に対して，1 文対毎に敬語への翻訳であるかどうかの情報を特殊トークンを付与することで，推論時に任意の敬意表現での翻訳を可能にした．本研究では，訂正度の情報を Sennrich ら [15] と同様に訓練データに特殊トークンとして付与することで，入力文に対する訂正度の制御を行う．

また，出力の多様化を行う研究についても文法誤り訂正の分野では行われていないが，他分野においては幾つか行われている．Shen ら [16] は，mixture of experts (MoE) モデルを機械翻訳に用いることで多様な出力の生成を行った．このモデルは，1 つの入力文に対して異なる expert を用いることで多様な出力の生成を可能にしている．本研究では，モデルの構造は変えず，訓練データのみを変更することにより多様化を目指した．Li ら [17] は，ニューラルネットワークを用いた対話モデルにおいて，応答文の多様化を行った．一般に，対話モデルの出力として汎用的な応答が生成されることが多いが，推論時にのみ文脈と応答の相互情報量を最大化するような応答を生成させることにより多様化を行った．Vijayakumar ら [12] は，推論時にビームを複数のグループに分け，グループ毎に順にビームサーチを行い，同じタイムステップ内の他のグループにて選択されたトークンに対して選択されにくくする制約を加えることで出力に多様性をもたらす手法を提案した．本研究では，同様に推論時にのみ訂正が必要な箇所に対して多様にする制約を加える手法を提案する．

2.3 事前学習済言語モデル

近年，自然言語処理の様々なタスクにおいて事前学習済言語モデルが利用されている．一般に，事前学習済言語モデルは大規模な学習データを用いることでタスクに依らない汎用的な言語表現を獲得した言語モデルであり，様々なタスクにおいて高い性能を発揮している．実際に，文法誤り訂正においても代表的な事前学習済モデルの一つである Bidirectional Encoder Representations from Transformers (BERT) [18] が用いられている [6]．BERT は Transformer の構造を利用しており，大規模なデータを用いて Masked Language Model (MLM) と Next Sentence Prediction (NSP) の 2 つの教師なし事前学習を行う．MLM では，学習データ内の一部のトークンを [MASK] トークンという特殊トークンに置換し，元々のトークンの推定を行う．NSP では，BERT に対して 2 文入力し，入力した 2 文が隣接した 2 文か否かの推定を行う．この 2 つの教師なし事前学習を同時に行うことで汎用的な言語表現を獲得している．また，BERT は，入力の先頭に [CLS] という特殊トークンを挿入し学習を行っている．その結果，入力文中のトークン毎の分散表現

だけでなく，[CLS] トークンの分散表現から入力文全体の分散表現を獲得することができる．さらに，BERT を用いて分類タスクを解く場合，事前学習済みの BERT の出力を入力とした分類器を用いて分類を行う．このとき，分類器の学習と同時に分類タスクのデータを用いて BERT の再学習も行う．本研究では，BERT を用いた文分類により，訂正度の推定や訂正箇所の分類を行った．

第 3 章　訂正度を制御した文法誤り訂正モデル

本研究では，訓練データより求められた単語編集率に基づいて特殊トークンの付与を行い，新たな訓練データに基づき文法誤り訂正モデルを訓練することで，推論時に特殊トークンによって訂正度の制御可能な文法誤り訂正モデルを提案する．図 3.1 に概要図を示す．図中の太字は訂正箇所を示している．

3.1　訓練方法

初めに，訓練データ内の誤りを含む誤り文と，それに対応する訂正が行われた訂正文から挿入の回数，削除の回数，置換の回数の和が最小となるように動的計画法を用いて編集距離を計算する．そして，求めた編集距離を誤り文の文長で割り，単語編集率を計算する．算出した単語編集率を基に訓練データをソートし，文数が均等になるように L 個の文集合に分割する．その分割した文集合ごとに異なる特殊トークンを定め， 誤り文の文頭に付与する．

この様にして，文頭に単語編集率によって定められた特殊トークンが付与された誤り文とそれに対応する訂正文の訓練データを作成した．この新たに作成した訓練データを用いて文法誤り訂正モデルを学習する．

3.2　推論方法

推論時，入力文の文頭に L 個の特殊トークンを付与することにより L 個の異なる訂正度にて訂正が行われた文が生成される．ここで，評価データの参照文から単語編集率を求めることはできないため，入力文に応じた適切な特殊トークンを選択する必要がある．しかし，1 つの入力文に対して訂正度はある程度定まるが，一意には定まらず，多様な訂正文を考慮した際に訂正度の幅が存在すると考えられる．図 3.2 は，CoNLL-2014 において，1 つの誤り文を訂正する際の訂正度にある程度の幅が存在することを示した図である．図中の点はある誤り文とそれに対応する人手で訂正を行った 1 つの参照文のペアを指しており，横軸は対象の誤り文に対する複数参照文の平均単語編集率，縦軸は対象のペアの単語編集率を表している．ま

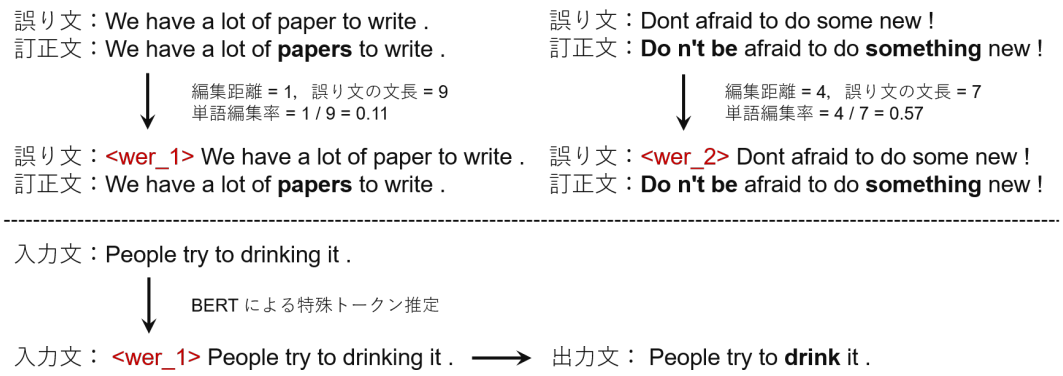

図 3.1: 訂正度を制御した文法誤り訂正モデルにおける訓練方法（上部）と推論方法（下部）の概要図

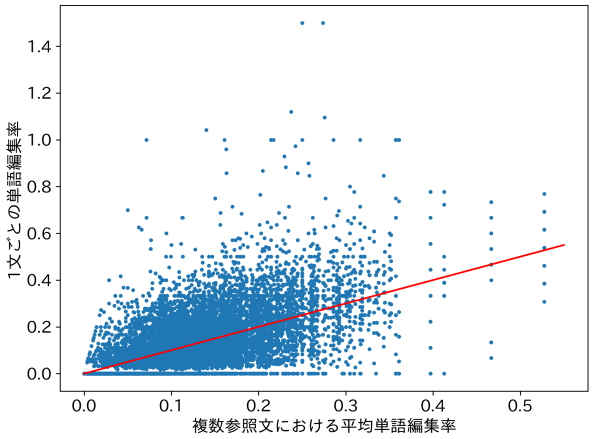

図 3.2: CoNLL-2014 の複数参照文における単語編集率の散布図

た，赤線は平均単語編集率を示した補助線である．この図より，全体的に複数参照文における平均単語編集率が高くなれば，参照文ごとの単語編集率も全体的に高くなるが，参照文ごとに単語編集率の幅が存在することがわかる．そこで，本研究では，BERT [18] を用いて入力文に対して訂正度の幅を考慮しながら適切な特殊トークンの選択を行った．詳細は 5.1 節にて説明する．

第 4 章　誤り箇所を考慮したビームサーチ

本研究では，訂正が必要な箇所のみに対して多様な訂正を行い，既に文法的に正しい箇所に対しては多様な訂正を行わない，誤り箇所を考慮したビームサーチを提案する．具体的には，誤り箇所を考慮したペナルティ $penalty$ を各タイムステップ t の各仮説 b 毎にモデルの対数出力確率に対して与えることで多様な出力の生成を目指した．ペナルティは以下の式で与えられる．

$$penalty_{b,t} = \lambda s_{b,t} + \beta \tag{4.01}$$

$s_{b,t}$ は，仮説 b, タイムステップ t における最も生成確率の高いトークンが，訂正が行われているトークンであるかどうかを [0,1] の範囲で表す指標である．具体的には，訂正が行われている場合は $s_{b,t}$ の値が 0 に近く，訂正が行われていない場合は $s_{b,t}$ の値が 1 に近くなることが望ましい．λ と β はハイパーパラメータであり，λ はペナルティの強さを調整し，β はペナルティがゼロになることを防ぐためのパラメータである．本研究では，BERT を用いて入力文とタイムステップ t までの出力文から訂正が行われているかどうかの分類を行い，訂正が行われていないと分類された確率を $s_{b,t}$ に用いた．詳細は，5.2 節にて説明する．以下の式のように，このペナルティを文法誤り訂正モデルの対数出力確率 $\log p$ に与えることで，ビームサーチのスコア k に対して制約を加えた．

$$k_{b,t} = penalty_{b,t} \log p_{b,t} \tag{4.02}$$

この制約により，訂正が行われていない仮説に対しては選択されにくくなり，反対に，訂正が行われている仮説に対しては選択されやすくなる．

図 4.1 は，入力文として “This essay will illustrate on two aspect .” が与えられた場合のビーム幅 2，タイムステップ t における概要図である．薄い色は生成確率が高いことを示している．図の上部は，直前のタイムステップ $t-1$ において “illustrate” を選択した第 1 仮説を，下部は，“this” を選択した第 2 仮説を表しており，色の薄さは生成確率の高さに比例している．このとき，タイムステップ t の第 1 仮説において最も生成確率の高いトークンは，入力文と同じ “on” であり，第 2 仮説においては，入力文から書き換えを行った “in” である．通常のビームサーチ

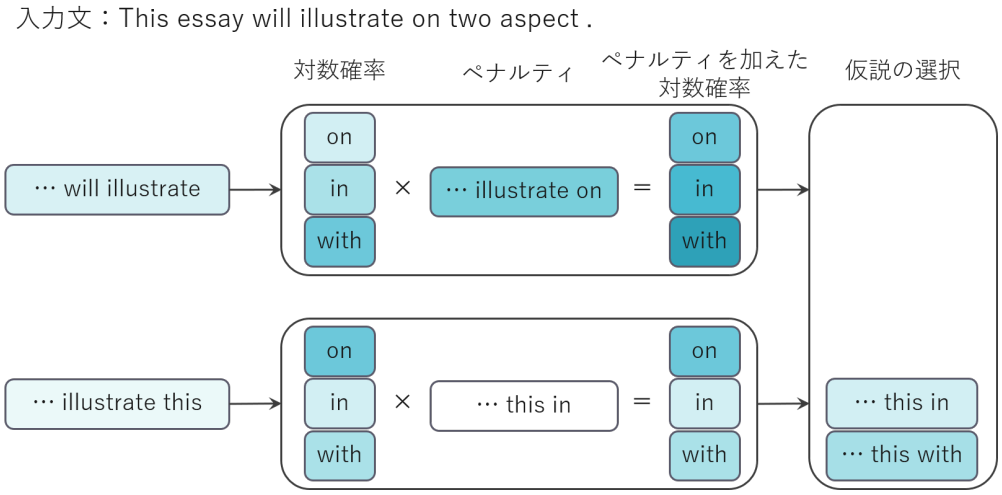

図 4.1: ビーム幅 2 における誤り箇所を考慮したビームサーチの概要図

では，それぞれ生成確率の高い，“illustrate on” と “this in” が選択される．しかし，本研究の提案手法では，第 1 仮説は書き換えを行っておらず，対して，第 2 仮説は書き換えを行っているため，第 1 仮説にのみペナルティが与えられ，最終的に訂正を行っている第 2 仮説から “this in” と “this with” が選択されることになる．

第 5 章　BERT を用いた文分類

BERT は大規模データにより事前学習された言語表現モデルであり，様々な自然言語処理のタスクにおいて高い性能を発揮している．本研究では，訂正度を制御した文法誤り訂正における適切な特殊トークンの選択や，誤り箇所を考慮したビームサーチにおける $s_{b,t}$ として BERT の [CLS] トークンの分散表現を分類する手法を用いた．いずれの手法も事前学習済みの BERT* に対して再訓練を行った．

5.1　BERT を用いた訂正度の推定

3 章にて述べた通り，訂正度を制御した文法誤り訂正モデルだけでは，参照文が与えられない限り，推論時に入力文に対して適切な訂正度の判別はできない．本章では，この BERT を用いた入力文に対して適切な訂正度の選択手法について述べる．

具体的な手法としては，BERT に対して誤り文のみを入力し，それに対応する訂正度を L 値分類として出力させるというものである．ここで，1 つの誤り文に対して複数の訂正度が考えられるが，誤り文に含まれる誤り箇所の個数は限られており，それに対応して単語編集率も限られているため，ある程度の分類は可能であると考えられる．

実際に訂正度を制御した文法誤り訂正モデルにおいて n-best 出力を得る際，BERT の予測確率を用いて複数の特殊トークンの出力から m_i-best を選択し，それらを合わせて最終的な n-best を得る．1 つの特殊トークンによる出力を用いるのではなく，複数の訂正度による出力を選択することで，より多様な出力を得ることができる．具体的な選択手法としては，まず，式 5.11 に示す通り， BERT の特殊トークンに対する出力スコア $o = (o_1, ..., o_i, ..., o_L)$ に対して Softmax 関数を用いて，特殊トークンごとの出力確率 $p = (p_1, ..., p_i, ..., p_L)$ を得る．得られた出力確率に対して式 5.12 に示す通り， n を掛け，小数点以下を四捨五入することで特殊トークン毎に採用する m_i-best の数を決定した．

*使用した事前学習済みモデル：`https://huggingface.co/bert-base-cased`

$$p_i = \frac{\exp(o_i)}{\sum_{l=1}^{L} \exp(o_l)} \tag{5.11}$$

$$m_i = \mathrm{round}(n \times p_i) \tag{5.12}$$

5.2 BERT による訂正箇所の分類

4 章で述べた通り，入力文とタイムステップ t までの出力文から訂正が行われているかどうかの分類を行う必要がある．動的計画法などを用いて入力文と出力文の対応を取ることで，訂正が行われているか否かの判定が可能になると考えられるが，データ内の文数 × ビーム幅 × タイムステップ数の処理が必要となる．本研究では，より高速にバッチ処理が可能な BERT を用いて入力文とタイムステップ t までの出力文から訂正が行われているかどうかの分類を行った．本節では，誤り箇所を考慮したビームサーチの $s_{b,t}$ に対する BERT の利用法について説明する．

$s_{b,t}$ は，4 章にて述べた通り，タイムステップ t，仮説 b の最も生成確率の高いトークンにおいて，訂正が行われているかどうかを表す指標である．そこで，BERT に対して誤り文と，その誤り文に対する，あるタイムステップまでのモデルの出力文の 2 文を入力し，モデルの出力文の最後のトークン，つまり，あるタイムステップ t，仮説 b において生成確率が最大であるトークンが訂正されたトークンであるかどうかの分類を行わさせる．そして，BERT が訂正していないと分類した確率を $s_{b,t}$ として用いる．

データセット	文数	参照文数	区分	評価手法
BEA-train	564,451	1	訓練	-
BEA-valid	4,384	1	開発	-
JFLEG	1,501	4	開発	GLEU
CoNLL-2014	1,312	18	評価	M^2

表 6.1: データセットの概要

第 6 章　実験

6.1　データセット

表 6.1 に本実験において使用したデータセットの概要を示す．訓練データとしては，学習者支援に関する国際会議である Workshop on Innovative Use of NLP for Building Educational Applications (BEA) において 2019 年に開催された Shared Task on Grammatical Error Correction [19] の Restricted track にて使用された訓練データ (BEA-train) を用いた．この訓練データには，The First Certificate in English（FCE）を受験した学習者の回答から集められた FCE コーパス [20]，言語学習者のための SNS である Lang-8 の添削データより集められた Lang-8 コーパス [7]，シンガポール国立大学の英語学習者である学生が作成した作文から集められた NUS Corpus of Learner English (NUCLE) [21]，英語学習者の学生のライティングを支援する Write & Improve と英語を母語とする学生の作文より集められた W & I+LOCNESS コーパス [19, 22] の 4 つのコーパスが含まれている．ただし，前処理として訂正が行われていない文と訂正結果が空白である文対は取り除いた．また，訓練時，最適なモデルのエポック数を決定するために，BEA-2019 において使用された開発データ (BEA-valid) を用い，この開発データに対する loss が最小となるエポック数を最適なモデルのエポック数とし，その時点でのモデルを実験に用いた．既存手法の多様なビームサーチ [12] や提案手法である誤り箇所を考慮したビームサーチに使用するハイパーパラメータに関しては，JFLEG を用い

て最適化を行った．JFLEG は 4 つの参照文が存在しており，CoNLL-2014 には 18 の参照文が存在する．ここで，CoNLL-2014 Shared Task: Grammatical Error Correction にて付与された参照文数は 2 つであるが，Bryant and Ng [1] により 8 つ，Sakaguchi ら [10] により 8 つの参照文が新たに作成されているため，それらを全て組み合わせ，18 の参照文数となる．

訂正度の推定に用いる BERT の再訓練には，訂正度を制御した文法誤り訂正モデルの訓練に用いた特殊トークンを付与したデータセットと同じ訓練データを用い，特殊トークンを付与する前の誤り文を入力し，そこから特殊トークンを予測させるという訓練を行った．ただし，ランダムに 5,000 文選択し，開発データとして用いた．

訂正箇所の分類に用いる BERT の再訓練に用いたデータセットとしては，NUCLE を用いた．ERRANT (grammatical ERRor ANnotation Toolkit) [23] を用いて，訂正文において誤り文から挿入・置換・削除の訂正が行われている箇所を抽出することで，誤り文と訂正が行われている単語までの訂正文のペアを作成した．訂正箇所の分類に用いる BERT においては訂正が行われていない場合，高いスコアであることが望まれるため，これを負例としてラベル付けした．ただし，削除の訂正に関しては，削除した単語の直後の単語までの訂正文を負例とした．正例としては，誤り文と訂正が行われていない単語までの訂正文のペアを作成し，負例と同数になるようにランダムに選択することで作成した．また，開発データとしては，CoNLL-2013 Shared Task on Grammatical Error Correction [24] にて用いられたデータセットである CoNLL-2013 に対して訓練データと同様の処理を行ったデータを用いた．以上の処理により，NUCLE は元々 57,151 文対存在するが，168,524 文対の学習データとなり，元々 1,381 文対存在する CoNLL-2013 は．11,009 文対の開発データとなった．また，再訓練の結果，開発データにおける F 値は約 0.97 であった．

6.2 評価方法

本研究では，手法毎の性能を比較するために以下の 3 つの評価指標を設定した．

6.2.1　訂正の多様性の評価（C-score）

訂正結果の多様性を評価するために，モデル出力の n-best と複数参照間の一致率を評価する．一致率の算出方法としては，機械翻訳における多様な翻訳文生成を目的とした 2020 Duolingo Shared Task [25] の評価指標として使用された重み付き再現率を使用する．具体的には，モデル出力の n-best と複数参照文間において完全一致した文数を求め，一致した参照文に応じて重み付けを行う．そして．重み付けされた値を評価データ内の全参照文数で割ることで一致率を算出する．ただし，本研究においては，ある 1 つの入力文に対する参照文間で重複した回数を参照文の総数で割った値を重みとして使用した．

6.2.2　訂正箇所の正しさの評価（DF-score）

入力文全体を書き換えて多様化するのではなく，実際に訂正が必要な箇所のみに書き換えが行われているかどうかの評価を行う．評価方法としては，まず式 6.21 に示す通り，誤り文 err の n-gram から参照文に対しての document frequency (df) を求める．ここで，$\boldsymbol{ng^{err}} = (ng_1^{err}, ..., ng_i^{err}, ..., ng_n^{err})$ は誤り文の n-gram 集合，$|\{r : r \ni ng_i^{err}\}|$ は ng_i^{err} を含む参照文 r の総数，$|R|$ は総参照文数である．次に，式 6.22 に示すとおり，誤り文の n-gram 集合とモデルの出力文 hyp の n-gram 集合 $\boldsymbol{ng^{hyp}}$ の積集合 $\boldsymbol{ng^{err \cap hyp}}$ を求める．そして，式 6.23 に示す通り，モデルの出力文における $ng_i^{err \cap hyp}$ の出現回数 $\mathrm{COUNT}(hyp, ng_i^{err \cap hyp})$ を求め，その n-gram に対応する df を出現回数に掛け，さらに誤り文における ng_i^{err} の出現回数 $\mathrm{COUNT}(err, ng_i^{err})$ に df を掛けた値の合計値で割ることで，そのモデル出力に対するスコアを算出する．

$$df_i = \frac{|\{r : r \ni ng_i^{err}\}|}{|R|} \tag{6.21}$$

$$\boldsymbol{ng^{err \cap hyp}} = \boldsymbol{ng^{err}} \cap \boldsymbol{ng^{hyp}} \tag{6.22}$$

$$df_{score} = \frac{\sum_{i=1}^{n} \mathrm{COUNT}(hyp, {ng_i}^{err \cap hyp}) df_i}{\sum_{i=1}^{n} \mathrm{COUNT}(err, ng_i^{err}) df_i} \tag{6.23}$$

このスコアは，df が高い n-gram，つまり書き換える必要のない n-gram がモデルにより書き換えられ，出力文に出現しない場合，大きく低下する．

6.2.3　文法誤り訂正の精度評価（G-score）

手法毎における文法誤り訂正タスクとしての性能を評価するため，一般的に文法誤り訂正タスクにおいて用いられる評価指標を用いて評価を行った．一般的に，JFLEG に対しては流暢な訂正が求められるため，単語の一致率などを考慮した GLEU score [26] を，CoNLL-2014 に対しては最小限の訂正のみであるため，訂正箇所のみの F 値で評価を行う MaxMatch (M^2) score [27] を用いて評価が行われる．そのため，本研究においても表 6.1 に示す通り，同様の評価尺度を用いて評価を行った．

6.3　文法誤り訂正モデル

本研究では，文法誤り訂正モデルとして Transformer base [14] を用い，パラメータも同様のものを用いた*．ベースラインモデルとしては，BEA-train をそのまま訓練データとして用いて訓練したモデルを用いた．ビームサーチ時のビーム幅は生成したい n-best 出力の n と同様の値を用いた．既存手法の多様なビームサーチ [12] におけるハイパーパラメータである Diversity Strength は 0.6 を，Number of Groups はビーム幅と同様の値を用いた．提案手法である誤り箇所を考慮したビームサーチのハイパーパラメータとして，ベースラインモデルに対しては $\lambda = 0.5$, $\beta = 1.0$ を，訂正度を制御した文法誤り訂正モデルに対しては $\lambda = 0.1$，$\beta = 1.0$ を用いた．

訂正度を制御した文法誤り訂正においては，BEA-train 内の全誤り文の文頭に対応する特殊トークンの付与を行った新たな訓練データを作成し，この新たに作成した訓練データを用いてモデルの訓練を行った．JFLEG において，異なる特殊トークンの種類数を用いて学習を行った結果を表 6.2 に示す．ここで，特殊トークンの種類数が 1 の場合は特殊トークンを付与していない通常の訓練データを用いた場合の結果を示している．BERT accuracy は BERT を用いた訂正度の推定精度を示しており，特殊トークンの種類数が多くなるにつれて推定精度が低下しており，その他の評価スコアも同様に低下している．したがって，BERT による推定精度

*使用したフレームワーク：`https://github.com/pytorch/fairseq/releases/tag/v0.9.0`

特殊トークン種類数	C-score	DF-score	G-score	BERT accuracy
1	29.13	85.14	**50.89**	-
2	**30.16**	**86.33**	50.41	**67.58**
4	29.75	86.23	49.81	43.33
6	29.18	86.16	49.91	35.14
8	29.01	85.77	49.21	31.24

表 6.2: JFLEG における特殊トークンの種類数別の結果

特殊トークン	最小値	最大値	文数
$\langle wer\_1 \rangle$	0.01	0.21	282,226
$\langle wer\_2 \rangle$	0.21	3.50	282,225

表 6.3: 訓練データ内の特殊トークンに対応する単語編集率の閾値と文数

が低いと適切でない訂正度の出力が選択されることになり，適切に多様性を向上できなかったと考えられる．特殊トークンの種類数としては 2 のときが最も高いスコアとなったため，本研究においては 2 を選択した．訂正度を制御した文法誤り訂正モデルの訓練に使用した特殊トークン毎の単語編集率の閾値を表 6.3 に示す．BEA-train 内の全文対における単語編集率に基づいて 2 つの集合に分け，$\langle wer\_1 \rangle$（単語編集率の低い文集合），$\langle wer\_2 \rangle$（単語編集率の高い文集合）の 2 つの特殊トークンを定義した．特殊トークン毎の単語編集率の最大値と最小値に重複が見られるのは，それぞれの特殊トークンの集合において文数が等しくなるように分割したためである．また，原文の単語数を超える単語数の挿入，削除，置換が行われる場合，単語編集率は 1 を超える．

Method	$n = 10$			$n = 15$			$n = 20$		
	C-score	DF-score	G-score	C-score	DF-score	G-score	C-score	DF-score	G-score
DGBS	20.60	89.03	48.50	21.30	83.63	48.35	21.76	81.46	48.40
MoE	18.08	**92.14**	43.10	18.56	83.19	43.57	16.54	95.34	44.45
PBS	28.40	87.65	48.56	29.58	86.44	**48.76**	30.41	85.67	**48.91**
+ DLBS	28.07	87.83*	**48.70**	29.72	86.79*	48.66	30.48	86.16*	48.52
WER	30.15*	89.77*	46.80	**31.55***	88.93*	46.80	32.52*	88.38*	46.70
+ DLBS	**30.22***	89.90*	46.86	31.53*	**89.08***†	46.80	**32.59***	**88.55***†	46.67

表 6.4: CoNLL-2014 における 10, 15, 20-best の実験結果

6.4 実験結果

表 6.4 は，ベースラインモデルに通常のビームサーチを適応したモデル (PBS)，多様なビームサーチを適応したモデル (DGBS) [12]，MoE [16] と，提案手法である訂正度を制御した文法誤り訂正モデル (WER)，WER に誤り箇所を考慮したビームサーチを適応したモデル (+DLBS) の評価データにおける実験結果である．また，10, 15, 20-best において実験を行った．アスタリスク（*）は提案手法が PBS に対して，ダガー（†）は +DLBS が WER に対して，ブートストラップ法により有意水準 0.05 で有意差があることを示す．

既存手法である DGBS と MoE は，PBS と比較すると出力の多様性を示す C-score が低下しており，機械翻訳などにおいては有効なこれらの手法が，文法誤り訂正においては有効でないことがわかる．さらに，15, 20-best において DGBS の DF-score が大きく低下していることから，訂正が不要な部分においても書き換えを行っていると考えられる．一方で，提案手法である WER は，G-score は低下しているが，C-score においていずれの既存手法よりも上回っており，DF-score も高いことから，WER は不要な訂正を行うことなく，多様な出力を生成できていることが確認できる．ここで，G-score の低下の主な原因として，“It is hereditary .” という入力文に対して，“It is a hereditary .” への訂正という不必要な冠詞の挿入などが多く見られ，多様化に伴い，細かい訂正が増加したことが挙げられる．DLBS を PBS に適用した場合は，10, 15-best において C-score が向上し，DF-score も

向上した．同様に，WER に適用した場合は 10, 20-best において C-score が向上し，DF-score も向上していることから訂正が必要な箇所のみを多様化していることが確認できる．

特殊トークン	JFLEG		CoNLL-2014			
	GLEU	単語編集率	P	R	$F_{0.5}$	単語編集率
$\langle wer\_1 \rangle$	45.66	0.12	**57.19**	24.00	44.80	0.04
$\langle wer\_2 \rangle$	49.40	0.17	49.97	34.44	**45.84**	0.08
$\langle wer\_3 \rangle$	**50.41**	0.23	42.71	39.25	41.97	0.13
$\langle wer\_4 \rangle$	47.88	0.32	38.01	**41.27**	38.62	0.21
Gold	-	0.24	-	-	-	0.15

表 7.1: 特殊トークン毎の誤り訂正結果

第 7 章　分析

7.1　訂正度の制御に関する分析

訂正度を制御した文法誤り訂正モデルにおいて，実際に特殊トークンによりモデルの訂正度が制御できているかどうかの実験を行った．異なる特殊トークンでの訂正度の違いを分析するため，この実験では特殊トークンの種類数を 4 として学習したモデルを利用した．そのため，この分析においては BEA-train 内の全文対における単語編集率に基づいて 4 つの集合に分け，$\langle wer\_1 \rangle$（単語編集率の最も低い文集合），$\langle wer\_2 \rangle$, $\langle wer\_3 \rangle$, $\langle wer\_4 \rangle$（単語編集率の最も高い文集合）の 4 つの特殊トークンを定義した．表 7.1 は，全て同一のモデルであるが，それぞれ推論時の入力文の全文に対して異なる特殊トークンを付与した場合の結果を示している．例えば，$\langle wer\_1 \rangle$ の行は JFLEG と CoNLL-2014 の全入力文の文頭に $\langle wer\_1 \rangle$ を付与して推論させた場合の結果である．

単語編集率の列は，それぞれの評価データの入力文とそれに対応するモデルの出力文から単語編集率の平均値を求めた結果であり，Gold の行は参照文との単語編集率の平均値である．これらの値を特殊トークンごとに比較すると，特殊トークンにより定義された単語編集率の大きさに比例して実際の単語編集率も変化していることがわかる．つまり，特殊トークンによってモデルの訂正度を制御することが可能であることを示している．しかし，同一特殊トークン内でも JFLEG と

CoNLL-2014 における単語編集率が異なることがわかる．一方で，同一特殊トークン内における単語編集率の差は Gold における単語編集率の差とほぼ同様であることから，このモデルは特殊トークンにより定義された単語編集率だけに従って訂正をするのではなく，入力文も考慮して訂正度を制御していると考えられる．

また，JFLEG において最も高い GLEU スコアとなったのは，特殊トークン $\langle wer\_3 \rangle$ を付与した場合であり，次は特殊トークン $\langle wer\_4 \rangle$ を付与した場合である．一方で，CoNLL-2014 において最も高い $F_{0.5}$ となったのは，特殊トークン $\langle wer\_2 \rangle$ を付与した場合であり，次は特殊トークン $\langle wer\_1 \rangle$ を付与した場合である．この差は，JFLEG と CoNLL-2014 において訂正度が異なるため，JFLEG においては CoNLL-2014 よりもより大きな訂正度で訂正を行った場合の方がスコアが向上したと考えられる．

それぞれの特殊トークンでの適合率（P）と再現率（R）を見ると，特殊トークン $\langle wer\_1 \rangle$ での適合率が最も高く，再現率が低い．一方で，特殊トークン $\langle wer\_4 \rangle$ での適合率が最も低く，再現率が高くなっている．このことから，単語編集率と比例して再現率が変化し，反対に単語編集率と反比例して適合率が変化していることがわかる．

7.2 BERT による特殊トークン推定に関する分析

図 7.1 は，CoNLL-2014 において，BERT による特殊トークン毎の予測結果と実際の参照文の単語編集率に基づいた特殊トークンとの混同行列を示している．縦軸は，CoNLL-2014 の複数参照文から算出した平均単語編集率に基づいた正解ラベルであり，横軸は，BERT により特殊トークン毎の m-best を選択する際の予測結果を示している．また，図中の値は実際に BERT により選択された m-best の総数を示しており，色の濃さは文数の多さと比例している．まず，表 7.1 に示した通り，CoNLL-2014 の平均単語編集率は低いため，正解ラベルも $\langle wer\_1 \rangle$ が多く，偏りが見られる．同様に，BERT による予測結果も $\langle wer\_1 \rangle$ と予測した結果が多くなっている．ただし，正解ラベルが $\langle wer\_2 \rangle$ の場合においても $\langle wer\_1 \rangle$ と予測した結果が多くなっているため，訂正度が低い方に予測を行う傾向があるということがわかる．

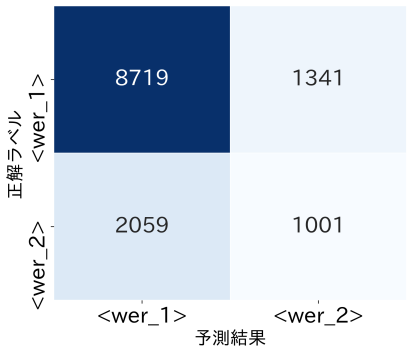

図 7.1: CoNLL-2014 における BERT による特殊トークンの推定結果

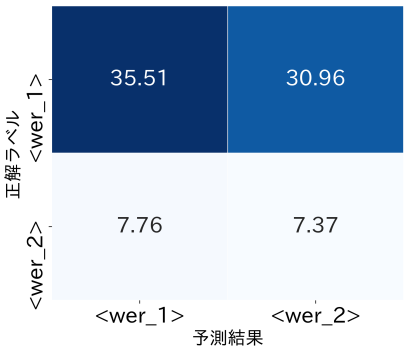

図 7.2: CoNLL-2014 における BERT により選択された特殊トークンごとの出力の C-score

図 7.2 は，正解ラベルと BERT による予測結果の組み合わせ毎の C-score を示した混同行列である．ただし，C-score を求める際，参照文を全文用いると BERT により出力文が選択されていない，つまり，出力文が存在しない場合，不当にスコアが低くなるため，出力文が選択されている入力文に対応する参照文のみを用いて評価した．スコアを比較すると，正解ラベルが $\langle wer\_1 \rangle$ であり，予測結果も $\langle wer\_1 \rangle$ の場合が最もスコアが高くなっている．また，正解ラベルが $\langle wer\_1 \rangle$ であり，予測結果が $\langle wer\_2 \rangle$ である文数は少ないが，スコアとしては高くなっている，これは，参照文に訂正が少ないデータが多いため，複数参照文内にも重複が多く，全体的にスコアが上がりやすいためであると考えられる．一方で，正解ラベルが $\langle wer\_2 \rangle$ の場合は，全体的にスコアが低くなっている．さらに，予測結果が $\langle wer\_2 \rangle$ である場合もスコアが高くなっておらず，予測結果が $\langle wer\_1 \rangle$ の場合のスコアの方が僅かに高くなっている．この結果の原因としては，まず，訂正度が高い方が訂正箇所が多く，文の完全一致で評価しているため正解ラベルが $\langle wer\_1 \rangle$ の場合よりも難しい問題となっているためであると考えられる．加えて，BERT により選択された文数が少なく，$\langle wer\_2 \rangle$ と予測された文数は $\langle wer\_1 \rangle$ と予測された文の半数であるため，スコアが低くなったと考えられる．

特殊トークン	$n = 10$		$n = 15$		$n = 20$	
	WER	+ DLBS	WER	+ DLBS	WER	+ DLBS
$\langle wer\_1 \rangle$	29.94	**30.06**	31.29	**31.39**	32.39	**32.40**
$\langle wer\_2 \rangle$	26.79	**26.92**	**28.12**	28.10	29.10	**29.43**
Single WER	29.79	29.92	31.09	31.14	32.18	32.26
Multi WER	30.15	30.22	31.55	31.53	32.52	32.59
Gold WER	29.60	29.70	30.83	30.95	31.98	32.03
Oracle WER	31.10	31.17	32.38	32.44	33.44	33.50

表 7.2: 特殊トークン毎の出力文の多様性

7.3 多様化に関する分析

表 7.2 の上部は，訂正度を制御した文法誤り訂正モデルにおける特殊トークン毎の C-score において，誤り箇所を考慮したビームサーチの影響を比較した表である．この表より，ほぼ全ての出力において，誤り箇所を考慮したビームサーチを用いることで多様性が向上することがわかる．$\langle wer\_1 \rangle$ と $\langle wer\_2 \rangle$ を比較すると，訂正度が高い $\langle wer\_2 \rangle$ の方がスコアが低くなっているが，これは CoNLL-2014 の訂正度が元々低いため，訂正度を高くしてもスコアの向上が見られなかったと考えられる．

表中の Single WER は BERT によって入力文から予測された特殊トークンの中で最も予測確率の高い単一の特殊トークンの出力文を選択した場合の結果である．一方で，Multi WER は BERT の予測確率から複数の特殊トークンの出力文を選択した場合の結果であり，表 6.4 の WER や +DLBS と同様の結果である．これらを比較すると，Single WER では，スコアが低下しているため，1 つの入力文に対して単一の訂正度ではなく，複数の訂正度を用いることで多様な訂正文が得られることがわかる．また，Gold WER は複数参照文から求めた平均の単語編集率から入力文に付与する特殊トークンを設定した場合で，Oracle WER は特殊トークン $\langle wer\_1 \rangle$ と $\langle wer\_2 \rangle$ の $2 \times n$-best の中からスコアが最も高くなるように n-best を選択した場合のスコアを表している．この結果より，参照文より求めた特殊トー

クンを付与することでは，最良の結果が得られないことがわかる．7.1 節で述べた通り，入力文も考慮して訂正度の制御が行われており，特殊トークンにより定義された単語編集率に基づいてモデルの訂正度が変化するわけではないため，高いスコアが得られなかったと考えられる．一方で，最良な n-best を選択することで，スコアが大幅に向上していることから，異なる特殊トークン間の出力に大きく多様性が存在していると考えられる．

第 8 章　おわりに

本研究では，文法誤り訂正において多様な訂正文を生成するための手法を提案した．訓練データに対して単語編集率に基づいた特殊トークンの付与を行い，新たに作成した訓練データで文法誤り訂正モデルを訓練することで出力文の訂正度の制御を可能にした．また，このモデルにより生成された訂正文が既存手法よりも多様な訂正文を生成可能であることも示した．さらに，誤り箇所を考慮したビームサーチを組み合わせることで，訂正が必要な箇所のみをより多様にした出力が得られる事を示した．

本研究では，文法誤り訂正における出力文の多様化を目的とした．将来的には，言い換え生成の様な入力文と出力文が大部分で共通するような別タスクにおける応用手法も検討したい．

発表リスト

筆頭発表

1. 甫立健悟, 金子正弘, 勝又智, 小町守. **文法誤り訂正における訂正度を考慮した多様な訂正文の生成.** 自然言語処理, 28 巻, 2 号. 採録決定.
2. Kengo Hotate, Masahiro Kaneko and Mamoru Komachi. **Generating Diverse Corrections with Local Beam Search for Grammatical Error Correction.** In the 28th International Conference on Computational Linguistics (COLING). December 9, 2020.
3. 甫立健悟, 金子正弘, 小町守. **Autoencoder を用いた頑健な文の分散表現生成の検討.** NLP 若手の会第 14 回シンポジウム. August 27, 2019.
4. Kengo Hotate, Masahiro Kaneko, Satoru Katsumata and Mamoru Komachi. **Controlling Grammatical Error Correction Using Word Edit Rate.** In the 56th Annual Meeting of the Association for Computational Linguistics Student Research Workshop (ACL SRW). July 30, 2019.

共著発表

1. 小山碧海, 甫立健悟, 金子正弘, 小町守. **文法誤り訂正における複数の擬似誤り生成モデルの比較.** NLP 若手の会第 15 回シンポジウム. September 23, 2020.
2. 今藤誠一郎, 甫立健悟, 平澤寅庄, 金子正弘, 小町守. **機械翻訳における非自己回帰モデルの複数言語の出力分析.** NLP 若手の会第 15 回シンポジウム. September 23, 2020.
3. Masahiro Kaneko, Kengo Hotate, Satoru Katsumata and Mamoru Komachi. **TMU Transformer System Using BERT for Re-ranking at BEA 2019 Grammatical Error Correction on Restricted Track.** In 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 14): Shared Task. August 2, 2019.

謝辞

本論文の作成にあたり，丁寧に指導をして下さった小町守准教授に深く感謝致します．また，学部時代からメンターとして指導して下さった勝又さん，金子さんには研究を進めていく上で大変多くの助言を頂きました．本当に感謝しております．Lang-8 のデータ使用にあたり，データを共有して頂きました株式会社 Lang-8 の喜洋洋様に感謝申し上げます．小町研究室に所属していた学生の皆さんには，研究以外の面においても大変お世話になり，楽しく研究生活を過ごすことができました．大変ありがとうございました．最後に，副査を引き受けてくださった山口亨教授と高間康史教授に感謝申し上げます．

参考文献

[1] C. Bryant and H.T. Ng, “How far are we from fully automatic high quality grammatical error correction?,” Proc. of ACL, pp.697–707, 2015.

[2] S. Chollampatt and H.T. Ng, “A multilayer convolutional encoder-decoder neural network for grammatical error correction,” Proc. of AAAI, pp.5755–5762, 2018.

[3] M. Junczys-Dowmunt, R. Grundkiewicz, S. Guha, and K. Heafield, “Approaching neural grammatical error correction as a low-resource machine translation task,” Proc. of NAACL, pp.595–606, 2018.

[4] W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu, “Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data,” Proc. of NAACL, pp.156–165, 2019.

[5] S. Kiyono, J. Suzuki, M. Mita, T. Mizumoto, and K. Inui, “An empirical study of incorporating pseudo data into grammatical error correction,” Proc. of EMNLP, pp.1236–1242, 2019.

[6] M. Kaneko, M. Mita, S. Kiyono, J. Suzuki, and K. Inui, “Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction,” Proc. of ACL, pp.4248–4254, 2020.

[7] 水本智也，小町　守，永田昌明，松本裕治，“日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得，” 人工知能学会論文誌，vol.28，no.5，pp.420–432，2013.

[8] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, “The CoNLL-2014 shared task on grammatical error correction,” Proc. of CoNLL, pp.1–14, 2014.

[9] C. Napoles, K. Sakaguchi, and J. Tetreault, “JFLEG: A fluency corpus and benchmark for grammatical error correction,” Proc. of EACL, pp.229–234, 2017.

[10] K. Sakaguchi, C. Napoles, M. Post, and J. Tetreault, “Reassessing the goals of grammatical error correction: Fluency instead of grammaticality,” TACL, vol.4, pp.169–182, 2016.

[11] R. Grundkiewicz, M. Junczys-Dowmunt, and K. Heafield, “Neural grammatical error correction systems with unsupervised pre-training on synthetic data,” Proc. of BEA, pp.252–263, 2019.

[12] A.K. Vijayakumar, M. Cogswell, R.R. Selvaraju, Q. Sun, S. Lee, D.J. Crandall, and D. Batra, “Diverse beam search for improved description of complex scenes,” Proc. of AAAI, pp.7371–7379, 2018.

[13] I. Kulikov, A. Miller, K. Cho, and J. Weston, “Importance of search and evaluation strategies in neural dialogue modeling,” Proc. of INLG, pp.76–87, 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," Proc. of NIPS, pp.5998–6008, 2017.

[15] R. Sennrich, B. Haddow, and A. Birch, "Controlling politeness in neural machine translation via side constraints," Proc. of NAACL, pp.35–40, 2016.

[16] T. Shen, M. Ott, M. Auli, and M. Ranzato, "Mixture models for diverse machine translation: Tricks of the trade," Proc. of ICML, pp.5719–5728, 2019.

[17] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," Proc. of NAACL, pp.110–119, 2016.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. of NAACL, pp.4171–4186, 2019.

[19] C. Bryant, M. Felice, Ø.E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction," Proc. of BEA, pp.52–75, 2019.

[20] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," Proc. of ACL, pp.180–189, 2011.

[21] D. Dahlmeier, H.T. Ng, and S.M. Wu, "Building a large annotated corpus of learner English: The NUS corpus of learner English," Proc. of BEA, pp.22–31, 2013.

[22] S. Granger, "The computer learner corpus: A versatile new source of data for SLA research.," Sylviane Granger, editor, Learner English on Computer, pp.3–18, 1998.

[23] C. Bryant, M. Felice, and T. Briscoe, "Automatic annotation and evaluation of error types for grammatical error correction," Proc. of ACL, pp.793–805, 2017.

[24] H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The CoNLL-2013 shared task on grammatical error correction," Proc. of CoNLL, pp.1–12, 2013.

[25] S. Mayhew, K. Bicknell, C. Brust, B. McDowell, W. Monroe, and B. Settles, "Simultaneous translation and paraphrase for language education," Proc. of WNGT, pp.232–243, 2020.

[26] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," Proc. of ACL, pp.588–593, 2015.

[27] D. Dahlmeier and H.T. Ng, "Better evaluation for grammatical error correction," Proc. of NAACL, pp.568–572, 2012.